palgrave communications
HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

https://doi.org/10.1057/s41599-019-0279-9 OPEN

Fake news game confers psychological resistance against online misinformation

Jon Roozenbeek[1] & Sander van der Linden[2]

ABSTRACT The spread of online misinformation poses serious challenges to societies worldwide. In a novel attempt to address this issue, we designed a psychological intervention in the form of an online browser game. In the game, players take on the role of a fake news producer and learn to master six documented techniques commonly used in the production of misinformation: polarisation, invoking emotions, spreading conspiracy theories, trolling people online, deflecting blame, and impersonating fake accounts. The game draws on an inoculation metaphor, where preemptively exposing, warning, and familiarising people with the strategies used in the production of fake news helps confer cognitive immunity when exposed to real misinformation. We conducted a large-scale evaluation of the game with $N = 15{,}000$ participants in a pre-post gameplay design. We provide initial evidence that people's ability to spot and resist misinformation improves after gameplay, irrespective of education, age, political ideology, and cognitive style.

[1] Department of Slavonic Studies and Department of Psychology, University of Cambridge, Cambridge, UK. [2] Department of Psychology, School of Biological Sciences, University of Cambridge, Downing Site, Cambridge CB2 3EB, UK. Correspondence and requests for materials should be addressed to S.v.d.L (email: sander.vanderlinden@psychol.cam.ac.uk)

Introduction

The rapid spread of "fake news" and online misinformation is a growing threat to the democratic process (Lewandowsky, Ecker, and Cook, 2017; van der Linden, et al., 2017a; Iyengar and Massey, 2018), and can have serious consequences for evidence-based decision making on a variety of societal issues, ranging from climate change and vaccinations to international relations (Poland and Spier, 2010; van der Linden, 2017; van der Linden et al., 2017b; Lazer et al., 2018). In some countries, the rapid spread of online misinformation is posing an additional, physical danger, sometimes leading to injury and even death. For example, false kidnapping rumours on WhatsApp have led to mob lynchings in India (BBC News, 2018a; Phartiyal, Patnaik, and Ingram, 2018).

Social media platforms have proven to be a particularly fertile breeding ground for online misinformation. For example, recent estimates suggest that about 47 million Twitter accounts (~15%) are bots (Varol et al., 2017). Some of these bots are used to spread political misinformation, especially during election campaigns. Recent examples of influential misinformation campaigns include the MacronLeaks during the French presidential elections in 2017 (Ferrara, 2017), the PizzaGate controversy during the 2016 U.S. Presidential elections, and rumours circulating in Sweden about the country's cooperation with NATO (Kragh and Åsberg, 2017).

A broad array of solutions have been proposed, ranging from making digital media literacy part of school curricula (Council of Europe, 2017; Select Committee on Communications, 2017), to the automated verification of rumours using machine learning algorithms (Vosoughi, Mohsenvand, and Roy, 2017) to conducting fact-checks in real-time (Bode and Vraga, 2015; Sethi, 2017). However, decades of research on human cognition finds that misinformation is not easily corrected. In particular, the continued influence effect of misinformation suggests that corrections are often ineffective as people continue to rely on debunked falsehoods (Nyhan and Reifler, 2010; Lewandowsky et al., 2012). Importantly, recent scholarship suggests that false news spreads faster and deeper than true information (Vosoughi, Roy, and Aral, 2018). Accordingly, developing better debunking and fact-checking tools is therefore unlikely to be sufficient to stem the flow of online misinformation (Chan et al., 2017; Lewandowsky, Ecker, and Cook, 2017).

In fact, the difficulties associated with "after-the-fact" approaches to combatting misinformation have prompted some researchers to explore preemptive ways of mitigating the problem (Cook, Lewandowsky, and Ecker, 2017; van der Linden et al., 2017b; Roozenbeek and van der Linden, 2018). The main thrust of this research is to prevent false narratives from taking root in memory in the first place, focusing specifically on the process of preemptive debunking or so-called "prebunking".

Originally pioneered by McGuire in the 1960s (McGuire and Papageorgis, 1961, 1962; McGuire, 1964; Compton, 2013), inoculation theory draws on a biological metaphor: just as injections containing a weakened dose of a virus can trigger antibodies in the immune system to confer resistance against future infection, the same can be achieved with information by cultivating mental antibodies against misinformation. In other words, by exposing people to a weakened version of a misleading argument, and by preemptively refuting this argument, attitudinal resistance can be conferred against future deception attempts. Meta-analytic research has found that inoculation messages are generally effective at conferring resistance against persuasion attempts (Banas and Rains, 2010).

Importantly, inoculation theory was developed well before the rise of the internet and traditionally, research has focused on protecting "cultural truisms", or beliefs so widely held that they are seldom questioned ("it's a good idea to brush your teeth", McGuire, 1964). In fact, the initial inoculation metaphor was mostly applied to situations in which people had supportive preexisting beliefs and attitudes toward an issue. Only recently have researchers begun to extend inoculation theory to more controversial issues where people are likely to hold vastly different and often polarised belief structures, for example in the context of climate change (van der Linden et al., 2017b), biotechnology (Wood, 2007), and conspiracy theories (Banas and Miller, 2013; Jolley and Douglas, 2017).

Crucially, this line of work finds that inoculation can still be effective even when applied to those individuals who have already been exposed to misinformation (Cook, Lewandowsky, and Ecker, 2017; Jolley and Douglas, 2017; van der Linden et al., 2017b). Conceptually, this approach is analogous to the emerging use of "therapeutic vaccines" administered to those who already have the disease. Therapeutic vaccines can bolster host defenses and still induce antiviral immunity (e.g., in the context of chronic infections and some cancers, see Autran et al., 2004). Similarly, those who already carry an informational "virus" can still benefit from inoculation treatments and become less susceptible to future persuasion and deception attempts. Recent advances in inoculation theory have called for both prophylactic and therapeutic tests of inoculation principles (Compton, 2019), which is especially relevant in the context of fake news and misinformation.

Yet, thus far, scholarship has primarily focused on inoculating study participants against persuasion attempts pertaining to a particular topic, such as climate change (van der Linden et al., 2017b) or 9/11 conspiracies (Banas and Miller, 2013). Although consistent with the initial theory, this approach presents fundamental problems in terms of both the scalability (Bonetto et al., 2018) and generalisability of the "vaccine" across issue domains (Roozenbeek and van der Linden, 2018). For example, recent work indicates that issuing a general warning before exposing participants to misinformation can offer a significant inoculation effect in itself (Bolsen and Druckman, 2015; Cook, Lewandowsky and Ecker, 2017; van der Linden et al., 2017b). This is consistent with a larger literature on the effectiveness of forewarnings and refutation in correcting misinformation (Ecker, Lewandowsky, and Tang, 2010; Walter and Murphy, 2018).

Importantly, by extending the interpretation of the immunisation metaphor, inoculation could provide a "broad-spectrum vaccine" against misinformation by focusing on the common tactics used in the production of misinformation rather than just the content of a specific persuasion attempt. For example, inoculation messages are known to spill-over to related but untreated attitudes, offering a "blanket of protection" (McGuire, 1964; Parker, Rains, and Ivanov, 2016). Moreover, recent research has provided some support for the idea that inoculation can emerge through exposing misleading arguments (Cook, Lewandowsky, and Ecker, 2017). Thus, we hypothesise that by exposing the general techniques that underlie many (political) persuasion attempts, broad-scale attitudinal resistance can be conferred. In considering how resistance is best promoted, it is important to note that prior research has primarily relied on providing passive (reading) rather than active (experiential) inoculations (Banas and Rains, 2010). In other words, participants are typically provided with the refutations to a certain misleading argument. However, as McGuire hypothesised in the 1960s (McGuire and Papageorgis, 1961), active refutation, where participants are prompted to actively generate pro- and counter-arguments themselves may be more effective, as internal arguing is a more involved cognitive process. This is relevant because inoculation can affect the structure of associative memory networks, increasing nodes and linkages between nodes (Pfau et al., 2005).

Building on this line of work, we are the first to implement the principle of active inoculation in an entirely novel experiential

Fig. 1 The "Bad News Game" intro screen (www.getbadnews.com)

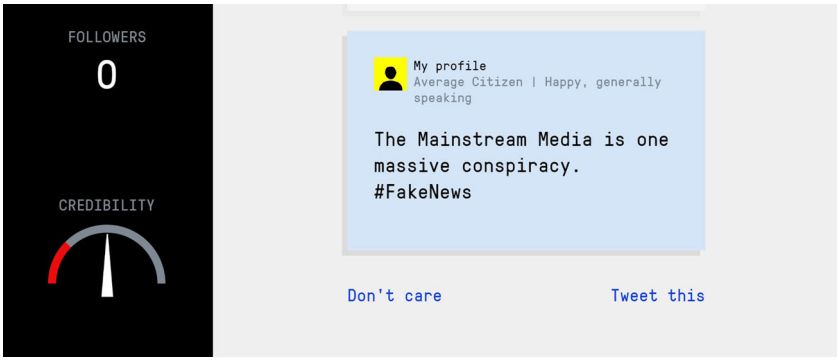

Fig. 2 Example tweet in the Bad News Game. *Note*: Follower and credibility metres are shown on the left-hand side

learning context: the Fake News Game, a "serious" social impact game that was designed to entertain, as well as educate. Previous work has shown that social impact games are capable of prompting behavioural change, for example, in the domain of health (Thompson et al., 2010). We posit that providing cognitive training on a general set of techniques within an interactive and simulated social media environment will help people apply these skills across a range of issue domains.

Accordingly, we developed a novel psychological intervention that aimed to confer cognitive resistance against fake news strategies. The intervention consisted of a freely accessible browser game that takes approximately 15 minutes to complete. The game, called Bad News, was developed in collaboration with the Dutch media platform DROG (DROG, 2018; BBC, 2018b). The game engine is capable of rendering text boxes, images, and Twitter posts to simulate the spread of online news and media. The game is choice-based: players are presented with various options that will affect their pathway throughout the game. Figure 1 shows a screenshot of the game's landing page.

In the game, players take on the role of a fake news creator. The purpose is to attract as many followers as possible while also maximising credibility. The follower and credibility metres along with a screenshot of the game environment are shown in Fig. 2.

Throughout the game, players gain followers and credibility by going through a number of scenarios, each focusing on one of six strategies commonly used in the spread of misinformation (NATO StratCom, 2017). At the end of each scenario, players earn a specific fake news badge (an overview of the fake news badges is provided in Fig. 3). Players are rewarded for making use of the strategies that they learn in the game, and are punished (in terms of losing credibility or followers) for choosing options in line with ethical journalistic behaviour. They gradually go from being an anonymous social media presence to running a (fictional) fake news empire. Players lose if their credibility drops to 0. The total number of followers at the end of the game counts as their final score.

There are six badges for players to earn, each reflecting a common misinformation strategy (see Fig. 3). The first badge is called "impersonation" and covers deception in the form of impersonating online accounts. This includes posing as a real person or organisation by mimicking their appearance, for example by using a slightly different username. This technique is commonly used on social media platforms, for example when impersonating celebrities, politicians, or in certain money and various other online scams (Goga, Venkatadri, and Gummadi, 2015; Jung, 2011; Reznik, 2013).

The second badge covers provocative emotional content: Producing material that deliberately plays into basic emotions such as fear, anger, or empathy, in order to gain attention or frame an issue in a particular way. Research shows that emotional content leads to higher engagement and is more likely to go viral and be remembered by news consumers (Aday, 2010; Gross and D'Ambrosio, 2004; Konijn, 2013; Zollo et al., 2015).

The third badge teaches players about group polarisation: Artificially amplifying existing grievances and tensions between different groups in society, for example political differences, in order to garner support for or antagonism towards partisan viewpoints and policies (Groenendyk, 2018; Iyengar and Krupenkin, 2018; Melki and Pickering, 2014; Prior, 2013).

The fourth badge lets players float their own conspiracy theories: Creating or amplifying alternative explanations for traditional news events which assume that these events are controlled by a small, usually malicious, secret elite group of people (Jolley and Douglas, 2017; Lewandowsky, Gignac, and Oberauer, 2013; van der Linden, 2015).

The fifth badge covers the process of discrediting opponents: Deflecting attention away from accusations of bias by attacking or delegitimising the source of the criticism (Rinnawi, 2007; Lischka,

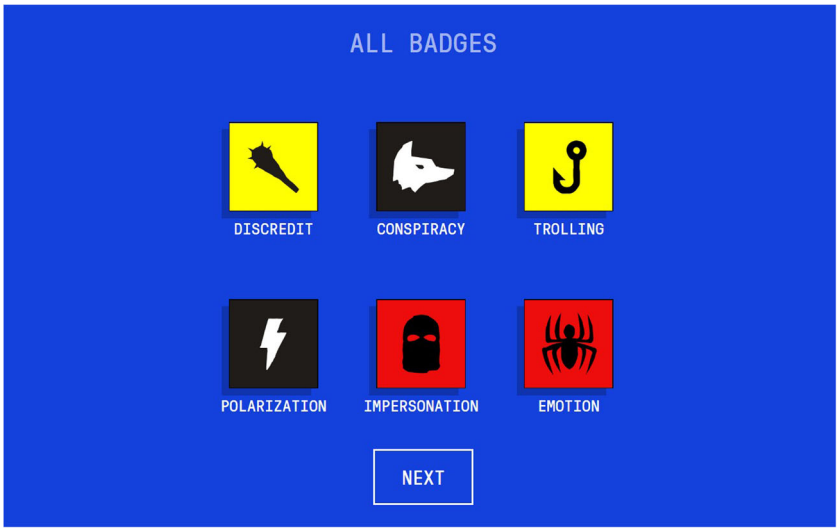

Fig. 3 The 6 badges that players earn throughout the game after successfully mastering a misinformation technique

2017), or by denying accusations of wrongdoing altogether (A'Beckett, 2013).

The last badge educates players about the practice of trolling people online. In its original meaning, the term trolling refers to slowly dragging a lure from the back of a fishing vessel in the hope that the fish will bite. In the context of misinformation, it means deliberately inciting a reaction from a target audience by using bait, making use of a variety of strategies from the earlier badges (Griffiths, 2014; McCosker, 2014; Thacker and Griffiths, 2012).

Methods

Sample and procedure. The game ('Bad News') is an interactive choice-based adventure (see Figs. 1–3). Players are shown a short text or image (such as a meme or an article headline) and can react to them in a variety of ways. Selecting an option that is in line with what a real producer of misinformation would choose earns players more followers and credibility. If, however, they lie too blatantly to their followers, choose an option that is overtly ridiculous, or act too much in line with journalistic best practices, the game either nudges followers onto a different path or lowers their credibility score.

As mentioned, during the approximately 15 minutes of playtime, players earn 6 badges by learning to apply six common misinformation techniques, namely: (1) impersonating people online (Goga, Venkatadri, and Gummadi, 2015; Jung, 2011; Reznik, 2013), (2) using emotional language (Aday, 2010; Gross and D'Ambrosio, 2004; Konijn, 2013; Zollo et al., 2015), (3) group polarisation (Groenendyk, 2018; Iyengar and Krupenkin, 2018; Melki and Pickering, 2014; Prior, 2013), (4) floating conspiracy theories (Jolley and Douglas, 2017; Lewandowsky, Gignac, and Oberauer, 2013; van der Linden, 2015) and building echo chambers (Flaxman, Goel, and Rao, 2016), (5) discrediting opponents (A'Beckett, 2013; Lischka, 2017; Rinnawi, 2007), and (6) trolling people online (Griffiths, 2014; McCosker, 2014; Thacker and Griffiths, 2012) and false amplification (NATO StratCom, 2017).

Participants were recruited through a press release with the university (the headline read; "*Fake news 'vaccine': online game may 'inoculate' by simulating propaganda tactics*). The release explained the research programme and provided an online web link to the game. The release was picked up by news outlets such as the BBC, which also provided a link to the game (BBC, 2018b). As such, we relied on a convenience sample in that anyone with an internet connection could visit the game website and participate. Using a traditional within-subjects design, the game featured a voluntary in-game (pre-post) survey for a period of three months. A few minutes into the game, we asked players if they wanted to participate in a scientific study. After players gave informed consent, we collected $N = 43{,}687$ responses over the three-month period following its launch (Feb–April 2018), which included $n = 14{,}266$ completed paired pre-post responses (the exact sample size and response rate differed slightly depending on the question, ranging from $n = 14{,}163$ to $n = 14{,}266$). All participants were automatically assigned a unique session ID and any duplicates were removed prior to analysis. The study was approved by the Cambridge Psychology Research Ethics Committee (PRE.2018.007). Socio-demographic variables were measured during the test, including gender (male, female, other), age (under 18, 19–29, 30–49, 50+), political orientation (measured on a 7-point scale where 1 is very left-wing and 7 is very right-wing), and highest education completed (high school or less, some college, higher degree). We also included the ball and bat-question from Frederick's cognitive reflection test (Frederick, 2005).

Given that the sample is self-selected and not representative of any particular population, the general distribution of the sample was skewed toward males (75%), higher educated (47%), younger (18–29, 47%), and somewhat-to-very-liberal (59%) individuals. Nonetheless, the sample size still allowed us to collect relatively large absolute numbers of respondents in each category (please see Suppl. Table S1 for full details on the sample and attrition).

Measures. The key dependent variable measured in the survey was respondents' ability to recognise misinformation strategies in the form of misleading tweets and news headlines. Participants were asked to rate the reliability of these tweets and headlines on a standard 7-point scale (1 = unreliable, 7 = reliable), both before and after playing. Because we did not want to overburden players with an excessively long survey, participants answered 6 questions in total (aside from the demographic questions), 2 of which were control questions that did not contain any deceptive techniques or strategies. Figure 4 shows a screenshot example of a survey question in the game module. The control questions were framed as tweets from legitimate news organisations that did not include any attempts at misleading the audience. The statements in these tweets were chosen to reflect global news events among English speakers. The first control question was a tweet by the New York Times, stating: *President Trump wants to build a wall between the*

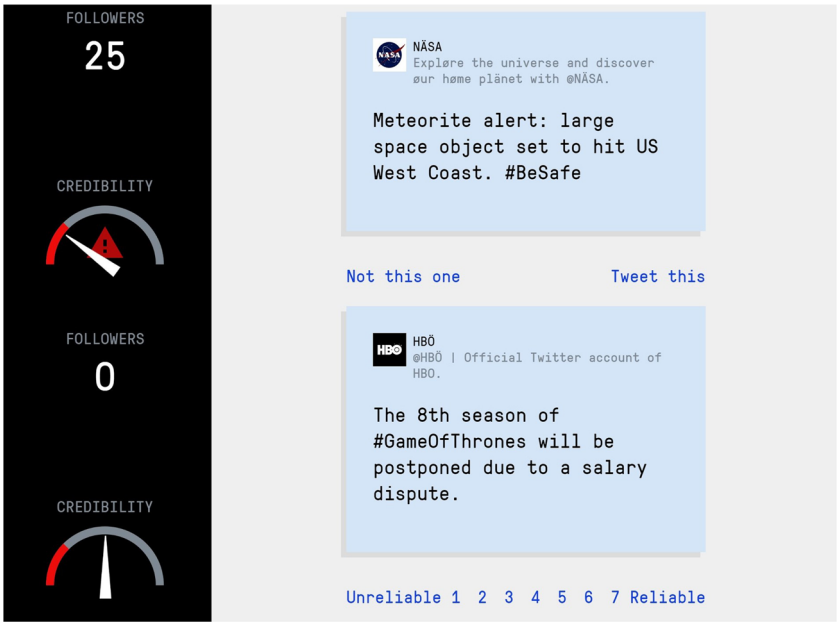

Fig. 4 Example of "training" and "testing" in the Bad News Game. *Note*: The top panel illustrates how a technique [impersonation] is used in the game, and the bottom panel shows how the same technique is used in a different example on which participants were evaluated before and after playing

United States and Mexico. The second was a tweet by the Wall Street Journal, stating *#Brexit, the United Kingdom's exit from the European Union, will officially happen in 2019*. The "real" items helped control for social desirability: if participants simply become more skeptical about all headlines they are shown, we would expect that they also rate the control items as significantly less reliable after playing.

Due to bandwidth and data storage limitations, the treatment questions reflected a random sample of the strategies included in the game: impersonation, conspiracy, and discrediting. Following the game's global popularity (BBC News, 2018b), additional data was gathered on polarisation. For impersonation, we used a well-known Twitter account with slight alterations in its username and avatar making a believable but untrue claim: participants were shown a tweet from an account impersonating the cable television company HBO, stating that *The 8th season of #GameOfThrones will be postponed due to a salary dispute*. This echoes a recent real-life example, where an impostor created a fake Twitter account imitating billionaire investor Warren Buffett. Although Buffett's name was misspelled as "Buffet", the account quickly gained a large following (BBC News, 2018c).

For the conspiracy question, participants were shown a tweet from a non-existent news site (Daily Web News) making a conspiratorial (Sunstein and Vermeule, 2009) claim: *The Bitcoin exchange rate is being manipulated by a small group of rich bankers. #InvestigateNow*. For the question about discrediting we used a different non-existent news site ("International Post Online"), employing an ad hominem (Walton, 1998) argument against the mainstream media: *The Mainstream Media has been caught in so many lies that it can't be trusted as a reliable news source. #FakeNews*. And finally, for polarisation, we showed participants an invented news headline that was randomised to state either that a *New study shows that left-wing people lie far more than right-wing people*, or the reverse (*New study shows that right-wing people lie far more than left-wing people*). We included this random element mainly to control for participants' political ideology. We hypothesised that people would rate each of the treatment (but not the control) items as less reliable after playing the game, thus displaying a cognitive inoculation effect.

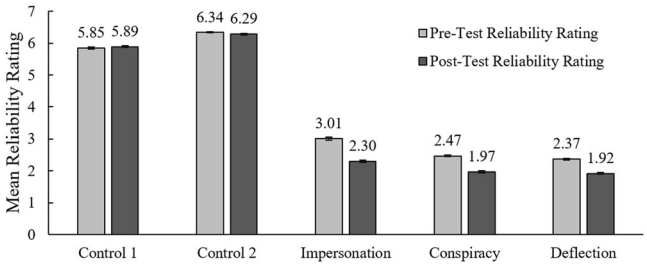

Fig. 5 Bar graph of pre (light grey) and post (dark grey) reliability judgments of control (real) and fake news items. *Note*: Error bars represent 95% confidence intervals

Results

The main results are displayed in Fig. 5. We analysed between $N = 14{,}163$ and $N = 14{,}266$ repeated within-subject measures from a survey embedded within the game. The exact number of pre-post completes varied slightly for each badge (see Supplementary Information for details). Participants rated the reliability of six headlines and tweets pre-and-post gameplay corresponding to some of the deception strategies that can be earned in the game: impersonation, floating conspiracies, discrediting opponents, political polarisation, as well as two "real" news control items (please also see "Methods").

Main effects of the intervention. A one-way repeated measures MANOVA on the five measures revealed a significant main effect, $F(5,\ 13559) = 980.65$, Wilk's $\Lambda = 0.73$, $p < 0.001$, $\eta^2 = 0.27$. We subsequently conducted a series of univariate follow-up comparisons with five paired *t*-tests using a conservative Bonferroni correction ($\alpha = 0.01$). However, given the large sample size, following Lakens (2013) we encourage the reader to also evaluate effect-sizes (Cohen's d_z, Hedges g_{av}) in addition to statistical significance. Furthermore, we provide violin plots visualising the full density distribution of the pre- and post-changes (Figs. 6 and 7).

Although statistically significant, there were no meaningful differences in the pre-scores and post-scores of the "real" control

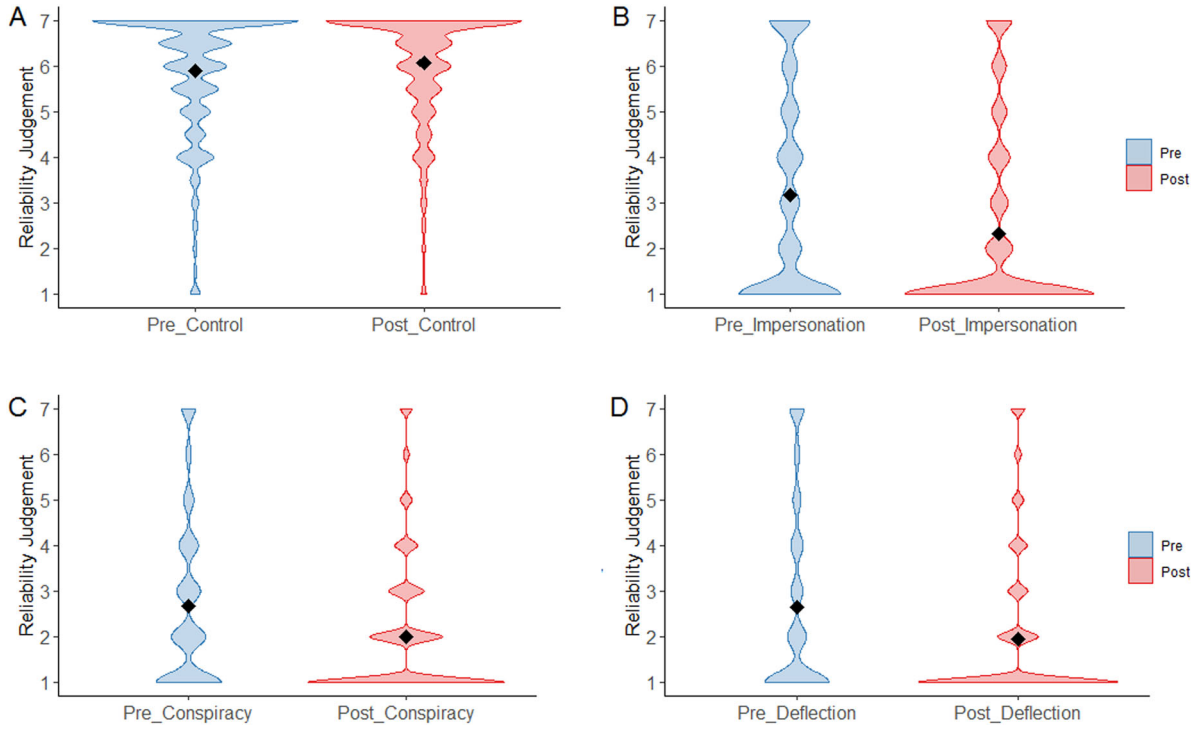

Fig. 6 Violin plots showing the kernel density distribution of pre-judgements and post-judgements with point estimates (block dots). *Note*: control "real" news items (collapsed, panel **a**), impersonation (panel **b**), conspiracy (panel **c**), and deflection (panel **d**)

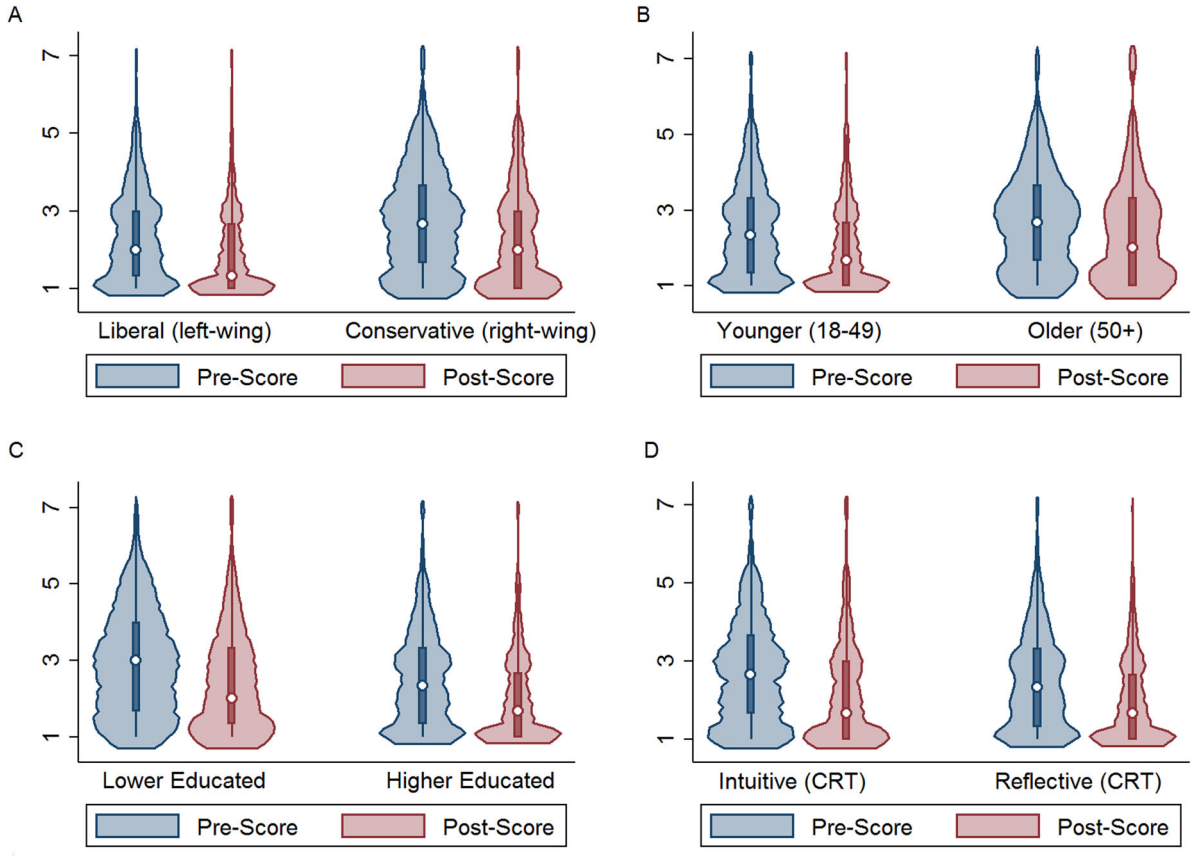

Fig. 7 Violin plots showing the kernel density distribution of pre-reliability and post-reliability judgements with median box plots (horizontal bars). *Note*: political ideology (panel **a**), age (panel **b**), education (panel **c**), and cognitive reflection (panel **d**)

headlines participants were presented with surrounding Donald Trump (control1), $M_{pre} = 5.85$, $M_{post} = 5.89$, $M_{diff} = 0.04$, [95% CI 0.02, 0.06], $t(14204) = 3.78$, $p = 0.002$, $d = 0.03$, Hedges $g = 0.02$ and Brexit (control2), $M_{pre} = 6.34$, $M_{post} = 6.29$, $M_{diff} = -0.05$, [95% CI −0.03, −0.07], $t(14265) = -5.13$, $p < 0.0001$, $d = 0.04$, Hedges $g = 0.04$. In contrast, there were both statistically significant and much larger differences in the pre-scores and post-scores for the fake tweets and headlines for each of the badges. Specifically, there was a significant decrease in reliability ratings for impersonation, $M_{pre} = 3.00$, $M_{post} = 2.30$, $M_{diff} = -0.71$, [95% CI −0.67, −0.74], $t(14224) = -43.42$, $p < 0.0001$, $d = 0.36$, Hedges $g = 0.33$, conspiracy, $M_{pre} = 2.47$, $M_{post} = 1.97$, $M_{diff} = -0.50$, [95% CI −0.48, −0.53], $t(14217) = -41.83$, $p < 0.0001$, $d = 0.35$, Hedges $g = 0.32$ and deflection, $M_{pre} = 2.37$, $M_{post} = 1.92$, $M_{diff} = -0.45$, [95% CI −0.43, −0.48], $t(14162) = -35.94$, $p < 0.0001$, $d = 0.30$, Hedges $g = 0.26$. Notably, although political polarisation is an important element of online misinformation, due to technical (storage) issues, data for the polarisation badge was not collected in the original launch. However, post-hoc data collection ($N = 1770$, $n = 885$ paired) allowed for an initial evaluation, which suggests a somewhat lower but still significant inoculation effect for polarising headlines, $M_{pre} = 2.57$, $M_{post} = 2.30$, $M_{diff} = -0.26$, [95% CI −0.15, −0.37], $t(854) = -4.76$, $p < 0.0001$, $d = 0.16$, Hedges $g = 0.15$. As a robustness check, we estimated all pre-post changes controlling for key sociodemographic covariates. Results remain virtually unchanged (see Suppl. Table S2).

Subgroup analyses. Because we did not have any strong a priori hypotheses to assume that individual differences would exist in performance across specific badges, we averaged across items for ease of interpretation when reporting the subgroup analyses that follow. Importantly, among those respondents who completed pre-post assessments for the included fake news badges, the average inoculation effect was relatively high, $M_{pre} = 2.61$, $M_{post} = 2.06$, $M_{diff} = -0.55$, [95% CI −0.53, −0.57], $t(13787) = -61.21$, $p < 0.001$, $d = 0.52$, Hedges $g = 0.43$. Moreover, when analysed by prior susceptibility, we find the largest effects among those individuals who were most likely to be vulnerable to fake news on the pre-test. In contrast, we find little evidence for meaningful variation across socio-demographics.

Political ideology. For example, although conservatives rated fake headlines more reliable than liberals at pre-test, $M_{cons} = 2.85$ vs. $M_{lib} = 2.38$, $M_{diff} = 0.47$, [95% CI 0.52, 0.42], $t(14032) = 19.66$, $p < 0.001$, $d = 0.17$, the average learning effect in post-pre scores between liberals and conservatives did not differ significantly (Fig. 7, panel A), $M_{\Delta lib} = -0.55$, $M_{\Delta cons} = -0.59$, $M_{diff} = -0.04$, [95% CI −0.08, 0.04], $t(1167) = -1.77$, $p = 0.08$, $d = 0.02$. This was also the case for the polarisation-item specifically, $M_{\Delta lib} = -0.23$, $M_{\Delta cons} = -0.32$, $M_{diff} = -0.09$, [95% CI −0.32, 0.15], $t(666) = -0.73$, $p = 0.46$, $d = 0.03$.

Age, education, gender, and cognitive reflection. There was a significant difference for age so that older players adjusted their reliability ratings somewhat less (Fig. 7, Panel B), although the standardised difference was so small that it can be considered negligible, $M_{\Delta younger} = -0.56$, $M_{\Delta older} = -0.41$, $M_{diff} = -0.15$, [95% CI −0.08, −0.22], $t(13666) = 4.37$, $p < 0.001$, $d = 0.04$. There was no significant difference across education levels (Fig. 7, panel C), $M_{\Delta lower} = -0.53$, $M_{\Delta higher} = -0.55$, $M_{diff} = -0.03$, [95% CI −0.09, 0.03], $t(13675) = -0.87$, $p = 0.38$, $d = 0.01$ nor across our single-item measure of cognitive reflection (Fig. 7, panel D), $M_{\Delta intuitive} = -0.56$, $M_{\Delta reflective} = -0.55$, $M_{diff} = -0.01$, [95% CI −0.02, 0.04], $t(13713) = -0.52$, $p = 0.60$, $d = 0.004$. Lastly, there was a significant gender difference so that females performed slightly better on average, but the effect-size was once again near negligible, $M_{\Delta female} = -0.66$, $M_{\Delta male} = -0.53$, $M_{diff} = -0.13$, [95% CI −0.09, −0.18], $t(13340) = 6.07$, $p < 0.001$, $d = 0.05$.

Prior susceptibility to fake news. In order to analyse the results by respondents' prior susceptibility to fake news, we created a median split based on how reliable people deemed the fake headlines to be at pre-test ($Mdn = 2.67$, $M = 2.82$, SD = 1.46). Findings reveal a much larger inoculation effect for those individuals who were more likely to think that the fake headlines were reliable at pre-test than those participants who were less susceptible, $M_{\Delta low} = -0.19$ vs. $M_{\Delta high} = -1.06$, $M_{diff} = -0.86$, [95% CI −0.83, −0.90], $t(13786) = 51.57$, $p < 0.001$, $d = 0.89$.

Discussion

We find preliminary evidence that the process of active inoculation through playing the Bad News game significantly reduced the perceived reliability of tweets that embedded several common online misinformation strategies. Although in absolute terms the standardised effect-sizes across the different badges may indicate a small to moderate effect, the observed magnitude is broadly in line with the average effect-size in the context of resistance to persuasion research (Banas and Rains, 2010; Walter and Murphy, 2018), where small effects are not only common but also meaningful, especially when aggregated across individuals over time (Funder and Ozer, 2019). For example, using a binomial effect-size display, consider that a Cohen's d of just 0.15 could roughly translate into a change of support for a particular policy by 7% (e.g., from 43% to 50%). Notably—while our effects were much larger—influential elections have been decided on substantially smaller margins (e.g., the EU Brexit referendum in 2016, 52% vs. 48%).

Moreover, rather than following the traditional inoculation approach where the "vaccine" is comprised of a weak dose of exactly the same (mis)information (*refutational-same*), the Bad News Game exposes participants to doses of weakened strategies and participants are subsequently tested using a range of different deception examples (*refutational-different*). The observation that participants rated both control questions as roughly equally reliable before and after playing the game underlines this point: it shows that active inoculation does not merely make participants more skeptical, but instead trains people to be more attuned to specific deception strategies. Or, to continue the metaphor, the psychological "vaccine" only activates specific "antibodies". Achieving this is much more challenging because participants are tasked with using active reasoning to recognise a range of common misinformation strategies in different contexts (rather than just retrieving facts from memory). In short, we consider these gains meaningful but encourage future research to further explore the boundary conditions of inoculation theory, for example, the extent to which the observed inoculation-effects extend beyond the game environment.

Furthermore, although we find some small variation in our results in line with previous research, such as that the elderly (Guess, Nagler, and Tucker, 2019) and conservatives (Grinberg et al., 2019; Guess, Nagler, and Tucker, 2019) may be more susceptible to fake news, we find no (practically) meaningful differences in inoculation-effects across genders, education levels, age groups, or political ideologies. This is a notable finding in itself, as our goal was to develop an intervention that could be used as a"broad-spectrum vaccine" without causing psychological reactance. This result is further buttressed by the finding that those participants who were most susceptible to fake news headlines at the outset (pre-test) also benefited the most from the

inoculation treatment. In other words, the vaccine may indeed help those audiences at greatest risk of misinformation. Of course, across the sample, the relatively modest susceptibility ratings indicate a possible floor effect among those who were already less likely to believe the fake headlines, as is common in fake news research (e.g., see Pennycook and Rand, 2018).

In addition, the uniform inoculation-effect across the political spectrum may be a result of the fact that we crafted each scenario to be ideologically balanced so that players always have an option to create fake news about a traditionally right-wing (anti-government) or left-wing (anti-industry) topic. Nevertheless, the observation that both liberals and conservatives improve in their ability to detect fake news following gameplay is broadly in line with the finding that susceptibility to fake news is at least partially explained by lack of appropriate reasoning skills rather than motivated cognition alone (Pennycook and Rand, 2018). Interestingly, prior research in this domain has often exposed participants to "real" fake news stories, which introduces a major memory confound (people may simply know whether a story is reliable or not because they remember it). Thus, one strength of the current approach is that the examples in the game are modelled after real instances of fake news—but fictional in nature—to rule out memory confounds.

Nonetheless, our study did suffer from a number of necessary limitations. First, during its official global launch with the university, it was not deemed ethical to design an intervention in which only some people would be randomised to play the Bad News game. In other words, just like it can be unethical to only assign some patients to a treatment, because the game is a social impact initiative that is supposed to be freely accessible to anyone, refusing half of the visitors the educational benefit would have been a disappointing experience, especially given the importance of the fake news debate. We thus lack a traditional control group and the results should be considered within the context of a non-randomised pre-post repeated measures design. We have made efforts to minimise this problem by including two "real news" control questions. If social desirability (or demand) effects were salient in the game, participants might have simply rated all items as less reliable, but this is not what we observed: participants did not meaningfully adjust their ratings of the "real" control items. Having said this, the control items were perhaps more likely to confound reliability with familiarity as their content covered major news events. Accordingly, alternative explanations cannot completely be ruled out, especially in the absence of a randomised control group. Yet, randomised between-subject studies in psychology often suffer from small samples mixed with high uncertainty around the estimates providing little useful information about the phenomenon of interest (Gelman and Carlin, 2014). Thus, there is something to be said in defense of the precision in our estimates as a result of leveraging a large online sample (Funder and Ozer, 2019).

Nonetheless, we recognise that the sample was self-selected (opt-in) and its composition is therefore unbalanced on several key demographics and not representative of any population (please see Supplementary Information Table S1 for full details). For example, the overrepresentation of males could be due to the fact that online gaming might stereotypically appeal more to men (Ivory, 2006). Unfortunately, due to the fact that the intervention was openly accessible for anyone to play, we could not include an extensive set of scales or multi-item survey questions, as this would have significantly interfered with people's willingness to play the game. Thus, we could not ascertain how confident people were in their assessments and so the scores may not reflect strongly held beliefs. Importantly, we also did not record any personally identifying information from participants (including IP addresses or location) to adhere to the latest European Union General Data Protection Regulations (GDPR). Yet we recognise the important possibility that participation rates may have been selectively higher among those audiences who would be more open to learning about fake news in the first place.

Lastly, the question can be raised as to whether we are encouraging people to use these insights to spread fake news or deceive other people online (i.e., a negative side effect). Although not empirically tested, we deem this risk extremely low for two reasons. First, while the game shows how easy it can be to start spreading deceptive content, the primary motivations for doing so are often political or financial in nature (Kirby, 2016; Gu, Kropotov and Yarochkin, 2017), neither of which are motivations elicited or provided by the game. In addition, ramping up an influential fake news machinery is very different from learning what deceptive content looks like. Second, none of the strategies and techniques shown in the game are secret (NATO StratCom, 2017); they are already being used to spread 'real' fake news, the game is simply helping people gain resistance against them.

Overall, despite these limitations, we highlight the potential of game-based psychological interventions to combat the problem of misinformation at the individual level. The participation rate and overall success of the game as a translational intervention (outside of the research context) further show that there is a high demand for evidence-based materials that help stem the flow of online misinformation. Lastly, the potential of psychological inoculation against fake news highlights the need to cultivate a "broad-spectrum" vaccine which targets a wide range of (evolving) misinformation pathogens. We offer initial evidence of spill-over effects or "blanket protection" against deception by focusing on the broader techniques and strategies that underpin the production of misinformation more generally rather than targeting only specific instances of fake news.

These preliminary findings open up many avenues for future research. For example, future studies could improve upon the current work by designing randomised controlled trials comparing the Bad News game against traditional media literacy tools (across different cultures), as well as model the rate of decay of the initial inoculation effect; this would yield important insights into whether and when "booster shots" (replay) may be required. In short, by designing and testing a novel experiential learning platform, we hope to have paved a path toward a new line of behavioural science research aimed at empowering diverse populations to guard themselves against the spread of misinformation in society.

Data availability

The data that support the findings of this study are available from https://doi.org/10.6084/m9.figshare.8269763.

Received: 16 April 2019 Accepted: 3 June 2019

References

A'Beckett L (2013) Strategies to discredit opponents: Russian representations of events in countries of the former Soviet Union. Psychol Lang Commun 17 (2):133–156. https://doi.org/10.2478/plc-2013-0009

Aday S (2010) Leading the charge: media, elites, and the use of emotion in stimulating rally effects in wartime. J Commun 60(3):440–465. https://doi.org/10.1111/j.1460-2466.2010.01489.x

Autran B, Carcelain G, Combadiere B, Debre P (2004) Therapeutic vaccines for chronic infections. Science 305(5681):205–208. https://doi.org/10.1126/science.1100600

Banas JA, Miller G (2013) Inducing resistance to conspiracy theory propaganda: testing inoculation and metainoculation strategies. Hum Commun Res 39(2):184–207

Banas JA, Rains SA (2010) A meta-analysis of research on inoculation theory. Commun Monogr 77(3):281–311. https://doi.org/10.1080/03637751003758193
BBC News (2018a) How WhatsApp helped turn an Indian village into a lynch mob. https://www.bbc.co.uk/news/world-asia-india-44856910
BBC News (2018b) Game helps players spot 'fake news https://www.bbc.co.uk/news/technology-43154667. Accessed 22 Feb 2018
BBC News (2018c) A fake billionaire is fooling people on Twitter. 28 August. https://www.bbc.co.uk/news/world-us-canada-45331781. Accessed 12 Dec 2018
NATO StratCom (2017) Digital hydra: security implications of false information online. https://www.stratcomcoe.org/digital-hydra-security-implications-false-information-online
Bode L, Vraga EK (2015) In related news, that was wrong: the correction of misinformation through related stories functionality in social media. J Commun 65(4):619–638. https://doi.org/10.1111/jcom.12166
Bolsen T, Druckman JN (2015) Counteracting the politicization of science. J Commun 65(5):745–769. https://doi.org/10.1111/jcom.12171
Bonetto E, Troian J, Varet F, Lo Monaco G, Girandola F (2018) Priming resistance to persuasion decreases adherence to conspiracy theories. Soc Influ 13 (3):125–136. https://doi.org/10.1080/15534510.2018.1471415
Chan MPS, Jones CR, Hall Jamieson K, Albarracin D (2017) Debunking: a meta-analysis of the psychological efficacy of messages countering misinformation. Psychol Sci 28(11):1531–1546. https://doi.org/10.1177/0956797617714579
Compton J (2013) Inoculation theory. In: Dillard JP, Shen L (eds) The SAGE Handbook of Persuasion: Developments in Theory and Practice. 2nd edn. SAGE Publications, Inc., Thousand Oaks. pp. 220–236
Compton, J (2019) Prophylactic versus therapeutic inoculation treatments for resistance to influence. Commun Theory. https://doi.org/10.1093/ct/qtz004
Cook J, Lewandowsky S, Ecker UKH (2017) 'Neutralizing misinformation through inoculation: exposing misleading argumentation techniques reduces their influence'. PLoS ONE 12(5):1–21. https://doi.org/10.1371/journal.pone.0175799
Council of Europe (2017) Media freedom, independence and diversity. www.coe.int. https://www.coe.int/en/web/commissioner/thematic-work/media-freedom?. Accessed 4 Sep 2017
DROG (2018) A good way to fight bad news. www.aboutbadnews.com. Accessed 20 Sep 2018
Ecker UKH, Lewandowsky S, Tang DTW (2010) Explicit warnings reduce but do not eliminate the continued influence of misinformation. Mem Cogn 38 (8):1087–1100. https://doi.org/10.3758/MC.38.8.1087
Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 French Presidential Election. First Monday. 22(8). https://doi.org/10.5210/fm.v22i8.8005
Flaxman S, Goel S, Rao JM (2016) Filter bubbles, echo chambers, and online news consumption. Public Opin Q 80(1):298–320. https://doi.org/10.1093/poq/nfw006
Frederick S (2005) Cognitive reflection and decision making. J Econ Perspect 19 (4):25–42. https://doi.org/10.1257/089533005775196732
Funder DC, Ozer DJ (2019). Evaluating effect size in psychological research: sense and nonsense. Adv Methods Prac Psychol Sci https://doi.org/10.1177/2515245919847202
Gelman A, Carlin J (2014) Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspect Psychol Sci 9(6):641–651
Goga O, Venkatadri G, Gummadi KP (2015) The Doppelgänger Bot Attack: Exploring Identity Impersonation in Online Social Networks. In: Proceedings of the 2015 Internet Measurement Conference. ACM, New York (IMC '15), pp. 141–153
Griffiths MD (2014) Adolescent trolling in online environments: a brief overview. Educ Health 32(3):85–87
Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D (2019) Fake news on twitter during the 2016 US Presidential election. Science 363 (6425):374–378. https://doi.org/10.1126/science.aau2706
Groenendyk E (2018) Competing motives in a polarized electorate: political responsiveness, identity defensiveness, and the rise of partisan antipathy. Political Psychol 39:159–171. https://doi.org/10.1111/pops.12481
Gross K, D'Ambrosio L (2004) Framing emotional response. Political Psychol 25 (1):1–29
Guess A, Nagler J, Tucker J (2019) Less than you think: prevalence and predictors of fake news dissemination on Facebook. Sci Adv 5(1). https://doi.org/10.1126/sciadv.aau4586
Gu L, Kropotov V, Yarochkin F (2017) The fake news machine: how propagandists abuse the internet and manipulate the public. TrendLabs Research Paper. https://documents.trendmicro.com/assets/white_papers/wp-fake-news-machine-how-propagandists-abuse-the-internet.pdf
Iyengar S, Krupenkin M (2018) The strengthening of Partisan affect. Political Psychol 39:201–218. https://doi.org/10.1111/pops.12487
Iyengar S, Massey DS (2018) Scientific communication in a post-truth society. Proc Natl Acad Sci. https://doi.org/10.1073/PNAS.1805868115
Ivory JD (2006) Still a man's game: gender representation in online reviews of video games. Mass Commun Soc 9(1):103–114
Jolley D, Douglas KM (2017) Prevention is better than cure: addressing anti-vaccine conspiracy theories. J Appl Soc Psychol https://doi.org/10.1111/jasp.12453
Jung AM (2011) Twittering away the right of publicity: personality rights and celebrity impersonation on social networking websites. Chic-Kent Law Rev 86 (1):381–418
Kirby EJ (2016, December 5) The city getting rich from fake news. *BBC News*. https://www.bbc.co.uk/news/magazine-38168281
Konijn EA (2013) The role of emotion in media use and effects. In: Dill KE (ed.), The Oxford Handbook of Media Psychology. New York/London: Oxford University press. pp. 186–211
Kragh M, Åsberg S (2017) Russia's strategy for influence through public diplomacy and active measures: the Swedish case. J Strateg Stud 40(6):773–816. https://doi.org/10.1080/01402390.2016.1273830
Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 4:863
Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M (2018) The science of fake news. Science 359(6380):1094–1096. https://doi.org/10.1126/science.aao2998
Lewandowsky S, Ecker UK, Seifert CM, Schwarz N, Cook J (2012) Misinformation and its correction: continued influence and successful debiasing. Psychol Sci Public Interest 13(3):106–131. https://doi.org/10.1177/1529100612451018
Lewandowsky S, Ecker UKH, Cook J (2017) Beyond misinformation: understanding and coping with the "post-truth" era. J Appl Res Mem Cogn 6 (4):353–369. https://doi.org/10.1016/j.jarmac.2017.07.008
Lewandowsky S, Gignac GE, Oberauer K (2013) The role of conspiracist ideation and worldviews in predicting rejection of science. PLOS ONE 8(10):1–11. https://doi.org/10.1371/journal.pone.0075637
Lischka JA (2017) A badge of honor?: how The New York Times discredits President Trump's fake news accusations. Journal Stud 1–18. https://doi.org/10.1080/1461670X.2017.1375385
McCosker A (2014) Trolling as provocation: YouTube's agonistic publics. Convergence 20(2):201–217. https://doi.org/10.1177/1354856513501413
McGuire WJ (1964) Inducing resistance against persuasion: some contemporary approaches. Adv Exp Soc Psychol 1:191–229. https://doi.org/10.1016/S0065-2601(08)60052-0
McGuire WJ, Papageorgis D (1961) Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. J Abnorm Soc Psychol 63:326–332
McGuire WJ, Papageorgis D (1962) Effectiveness of forewarning in developing resistance to persuasion. Public Opin Q 26(1):24–34. https://doi.org/10.1086/267068
Melki M, Pickering A (2014) Ideological polarization and the media. Econ Lett 125 (1):36–39. https://doi.org/10.1016/j.econlet.2014.08.008
Nyhan B, Reifler J (2010) When corrections fail: the persistence of political misperceptions. Political Behav 32(2):303–330. https://doi.org/10.1007/s11109-010-9112-2
Parker KA, Rains SA, Ivanov B (2016) Examining the "blanket of protection" conferred by inoculation: the effects of inoculation messages on the cross-protection of related attitudes. Commun Monogr 83(1):49–68. https://doi.org/10.1080/03637751.2015.1030681
Pennycook G, Rand DG (2018) Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition https://doi.org/10.1016/j.cognition.2018.06.011
Pfau M, Ivanov B, Houston B, Haigh M, Sims J, Gilchrist E, Russell J, Wigley S, Eckstein J, Richert N (2005) 'Inoculation and mental processing: the instrumental role of associative networks in the process of resistance to counterattitudinal influence'. Commun Monogr 72(4):414–441. https://doi.org/10.1080/03637750500322578
Phartiyal S, Patnaik S, Ingram D (2018) When a text can trigger a lynching: WhatsApp struggles with incendiary messages in India. *Reuters UK*, 25 June. https://uk.reuters.com/article/us-facebook-india-whatsapp-fake-news/when-a-text-can-trigger-a-lynching-whatsapp-struggles-with-incendiary-messages-in-india-idUKKBN1JL0OW
Poland GA, Spier R (2010) Fear, misinformation, and innumerates: how the Wakefield paper, the press, and advocacy groups damaged the public health. Vaccine 28(12):2361
Prior M (2013) Media and political polarization. Annu Rev Political Sci 16 (1):101–127. https://doi.org/10.1146/annurev-polisci-100711-135242
Reznik, M (2013) Identity theft on social networking sites: developing issues of internet impersonation. Touro Law Rev 29(2):455–484
Rinnawi K (2007) De-legitimization of media mechanisms: israeli press coverage of the Al Aqsa Intifada. Int Commun Gazette 69(2):149–178. https://doi.org/10.1177/1748048507074927

Roozenbeek J, van der Linden S (2018) The fake news game: actively inoculating against the risk of misinformation. J Risk Res 22(5):570–580. https://doi.org/10.1080/13669877.2018.1443491
Select Committee on Communications (2017) Growing up with the Internet. 130. https://publications.parliament.uk/pa/ld201617/ldselect/ldcomuni/130/13002.htm
Sethi, RJ (2017) Crowdsourcing the Verification of Fake News and Alternative Facts. In Proceedings of the 28th ACM Conference on Hypertext and Social Media. ACM, New York (HT '17), pp. 315–316
Sunstein CR, Vermeule A (2009) Conspiracy theories: causes and cures. J Political Philos 17(2):202–227. https://doi.org/10.1111/j.1467-9760.2008.00325.x
Thacker S, Griffiths MD (2012) An exploratory study of trolling in online video gaming. Int J Cyber Behav Psychol Learn 2(4):17–33. https://doi.org/10.4018/ijcbpl.2012100102
Thompson D, Baranowski T, Buday R, Baranowski J, Thompson V, Jago R, Griffith MJ (2010) Serious video games for health: How behavioral science guided the development of a serious video game. Simul gaming 41(4):587–606. https://doi.org/10.1177/1046878108328087
van der Linden S (2015) The conspiracy-effect: exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. Personal Individ Differ 87:171–173. https://doi.org/10.1016/j.paid.2015.07.045
van der Linden S (2017) Beating the hell out of fake news. Ethical Record 122(6):4–7
van der Linden S, Maibach E, Cook J, Leiserowitz A, Lewandowsky S (2017a) Inoculating against misinformation. Science 358(6367):1141–1142. https://doi.org/10.1126/science.aar4533
van der Linden S, Leiserowitz A, Rosenthal S, Maibach E (2017b) Inoculating the public against misinformation about climate change. Global Challenges 1 (2):1600008. https://doi.org/10.1002/gch2.201600008
Varol O, Ferrara E, Davis CA, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. In: Eleventh International AAAI conference on Web and Social Media. http://arxiv.org/abs/1703.03107
Vosoughi S, Mohsenvand M, Roy D (2017) Rumor gauge: predicting the veracity of rumors on twitter. ACM Transactions on Knowledge Discovery from Data 11 (4):50. https://doi.org/10.1145/3070644
Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359(6380):1146–1151. https://doi.org/10.1126/science.aap9559
Walter N, Murphy ST (2018) How to unring the bell: a meta-analytic approach to correction of misinformation. Commun Monogr 85(3):423–441. https://doi.org/10.1080/03637751.2018.1467564
Walton D (1998) Ad Hominem Arguments. The University of Alabama Press, Tuscaloosa and London
Wood MLM (2007) Rethinking the inoculation analogy: effects on subjects with differing preexisting attitudes. Hum Commun Res 33(3):357–378. https://doi.org/10.1111/j.1468-2958.2007.00303.x
Zollo F, Novak PK, Del Vicario M, Bessi A, Mozetič I, Scala A, Caldarelli G, Quattrociocchi W (2015) Emotional dynamics in the age of misinformation. PLoS ONE 10(9):e0138740. https://doi.org/10.1371/journal.pone.0138740

Acknowledgements

We thank the Economic and Social Research Council (ESRC) for funding this research, as well as DROG and the University of Cambridge. We are also grateful to statistician Breanne Chryst for her advice and help with the data, formatting, and coding.

Additional information

The online version of this article (https://doi.org/10.1057/s41599-019-0279-9) contains supplementary material, which is available to authorized users.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://www.nature.com/reprints

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019